Modelling of Indian Monsoon Rainfall Series by Univariate Box—Jenkins Type of Models①

S.D.Dahale and *S.V.Singh*
Indian Institute of Tropical Meteorology, Pune-411008, India
Received January 22, 1992; revised June 4, 1992

ABSTRACT

The time domain approach, i.e. Autoregressive (AR) processes, of time series analysis is applied to the monsoon rainfall series of India and its two major regions, viz. North-West India and Central India. Since the original time series shows no modelable structure due to the presence of high interannual variability, a 3-point running filter is applied before exploring and fitting appropriate stochastic models. Out of several parsimonious models fitted, AR(3) is found to be most suitable. The usefulness of this fitted model is validted on an independent datum of 18 years and some skill has been noted. These models therefore can be used for low skill higher lead time forecasts of monsoon. Further the forecasts produced through such models can be combined with other forecasts to increase the skill of monsoon forecasts.

I. INTRODUCTION

The Indian monsoon rainfall shows large variability with proportionate impact on Indian economy. This variability consists of two components, viz. (i) The well recognized interannual variability, which is the major contributor towards the total variability of Indian monsoon and (ii) The more subtle longer period epochal variability with periods of about 60–70 years (Pant et al., 1988). Because of its larger contribution, the interannual variability has been extensively studied and methods of forecasting developed by the researchers in the past. Recently Gowariker et al. (1989) (See also Thapliyal, 1990) have used 15 global / regional parameters to predict Indian monsoon. These studies are based on physical linkage between the predictors and the predictand (viz. Indian monsoon). However, as noted by Normand (1953), the Indian monsoon rainfall has its connection with posterior events rather than its earlier events. Therefore, it stands out as an active, not as a passive feature of global climate system more effective as forecasting tool than as an event to be forecasted. For example, Elliot and Angell (1987) have suggested that the Southern Oscillation Index (SOI) and Sea Surface Temperature (SST) in eastern equatorial Pacific can be anticipated by Indian Monsoon. Large interannual changes in the release of energy over India and adjacent regions could well affect subsequent circulation suggesting that the monsoon process may be physically linked with itself through feedback mechanism; such feedback process may control the evolutionary process of many dynamic series. The autocoherence indicating the intrinsic predictability has been examined for many climatic elements like Palmer Drought Index (Katz and Skaggs, 1981), Sunspots, Baltic ice and zonal circulation (Jolliffe, 1983), Rainfall (Yao, 1983), Southern Oscillation (Chu and Katz, 1985), SST (Brown and Flueck, 1987) and

①Part of this study is published in LRF Report No.14, Programme on Long-Range Forecasting Research, WMO / TD No.395 (1991), pp. 67–72.

area Indian rainfall (Thapliyal, 1990). The memory of about a decade in monsoon rainfall has been indicated by Thapliyal (1986). The interannual variability of monsoon rainfall has been modelled as ARIMA as well as Ex-ARIMA process by Thapliyal (1982, 1986). However, the autocorrelation function of all India monsoon time series exhibits no structure. It practically behaves like white noise or random process except that a barely significant autocorrelation at lag 14 is noted. Therefore it cannot be modelld by the ARIMA (Box-Jenkins type) class of models with any fruitful result. We however feel that the smoothed rainfall series representing longer period variability can be modelled by this class of models. For separating the long period cycle from the interannual variability, we use 3-term running average of the rainfall series.

In Section II we describe the data used. The relevant Univariate Box-Jenkins (UBJ) methodology is introduced in brief in Section III and the Kresults are preseted in Section IV.Finally the results are discussed and broad conclusions brought out in Section V.

II. DATA

The monsoon rainfall data of whole India (henceforth referred to as India) have been taken from Parthasarathy et al. (1990) for the period 1871-1989. The data for North West (N. W.) India and central India have been computed from data of meteorological subdivisions published by Parthasarathy et al. (1987) for the period 1871-1984. The data of these regions have been updated till 1987 through personal communication with Dr. B. Parthasarathy. Area of study and sample statistics based on 1871-1970 of these data are shown in Figure 1.

For examining the effect of smoothening on the original time series, the raw Indian monsoon rainfall series is smoothed by 3-, 5-, 7-, 9-, 11- and 13-point running mean filters. Smoothening by these filters retains 28%, 18%, 13%, 11%, 9%, and 7% of the original variance respectively. We use here 3-point runing mean values for modelling as it retains substantive (28%) amount of variance of the original time series. This filter eliminates the

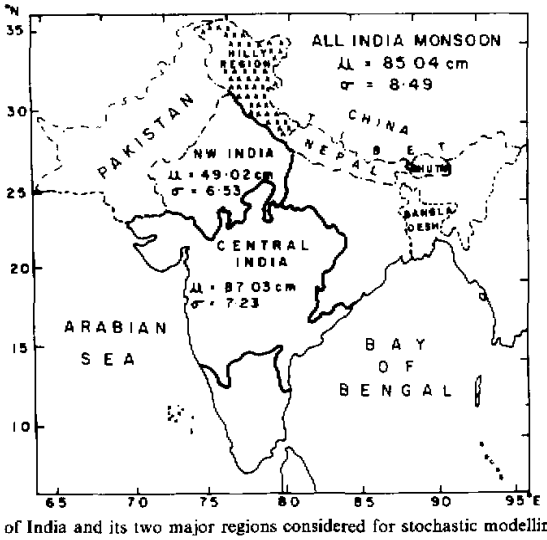

Fig.1. Map of India and its two major regions considered for stochastic modelling of monsoon rainfall.

influence of short period processes like stratospheric QBO, but retains the longer period cycles like Southern Oscillation, Sunspots etc., intended to be modelled.

III. METHODOLOGY IN BRIEF

1. *Univariate Box–Jenkins Approach*

The parametric time series model building constitutes an attempt to construct from a given set of data the underlying stochastic process that might have generated the realization. A general class of linear process is denoted by ARIMA (p, d, q, P, D, Q, s), where the small letters p, q refer to autoregressive and moving average parameters, d being the order of differencing for the non–seasonal component of the process whereas capitals refer to corresponding parameters / operations to the seasonal component. s is a measure of seasonality. One is generally interested to find the class of parsimonious model which describes given observations satisfactorily. However, before building appropriate model the basic conditions of normality and stationarity should be satisfied. As we consider only the monsoon, the question of seasonality does not come into picture here. Also instead of differencing we have used some smoothening to focus on certain desired aspects of the series. Therefore, the present description and the approach will generally be limited to ARMA (p, q) class of models.

2. *Model Building*

Let Z_t be a series of some stochastic variable z, available at equally spaced intervals of time. For non–zero mean (μ), an autoregressive moving average process is defined as:

$$Z_t = \mu + \sum_{j=1}^{p} \varphi_j (Z_{t-j} - \mu) + a_t - \sum_{k=1}^{q} \theta_k a_{t-k} \ ,$$

where φ_j are p autoregressive and Q_k, q moving average parameters and $a_t \sim N(0, \sigma^2 a)$ are normal random shocks or innovations. μ is a parameter that describes the level of process about which it fluctuates.

Box–Jenkins (1976) three stage strategy consisting of iterative cycle of identification, estimation and diagnostic checking (also see Pankratz, 1983) is followed in this study. We proceed on our discussions from final models identified. At the estimation and diagnostic stages the parameters are tested for thier quality such as statistical significance, closeness of fit [RMSE and Theil's U (Theil, 1966)] and intercorrelations (yielding joint sampling distributions of coefficients). Stationarity of process is checked through the admissibility of parameters.

3. *Diagnostic Checking*

At this stage the identified models are tested for adequacy by analyzing residuals which should resemble white noise. Two tests described in the following paragraphs are used.

a. Ljung–Box test

$$Q^* = n(n+2) \sum_{k=1}^{m} (n-k)^{-1} \hat{r}_k^2 (a) \ ,$$

where $\hat{r}_k (a)$ are autocorrelations of the residuals. This statistic is preferred to Box–Pierce Q since its sampling distribution more nearly approximates chi–squared when sample size is moderate (Pankratz, 1983).

b. Theil's test

Theil's (Theil, 1966) U-statistic is another measure of association between observed and modelled series. It compares how values (in sequence) in modelled series are changing in the light of change in the successive observed values. The smaller the value of U, the better the model. U is close to zero for most appropriate model.

4. *Model Validation*

The diagnostic tests carried out during the model building stage are not adequate to check the sufficiency of a stochastic model. The real test of a model lies in its performance in forecasting on new data. It is therefore generally advisable to test (or cross-validate) the model on an independent sample of reasonable size. Large number of simple statistics like Root Mean Square error, mean absolute error or correlation coefficient have been suggested and used for this purpose. Another commonly used score is the Heidke Skill Score (H. S. S.) which compares the skill scores of a forecast procedure against a naive forecast method, viz., climatology, random or persistence.

For computation of this score, the forecast and observed values of the variable are categorized in few classes (generally 3 for rainfall, viz. excess, normal and deficient) and contingency tables prepared. The skilll scores are then computed by using the following formula

$$S.S. = \frac{C - E}{T - E} ,$$

where C is the correct number of forecasts, E is the correct number of forecasts expected by naive method and T=Total number of forecasts. In the present study the normal values (middle class) are assumed to lie in the range between mean $+0.5\sigma$ and mean -0.5σ and excess (deficient) values lie above (below) the upper (lower) limit of normal (σ = standard deviation). The SS is zero when the numbers of correct forecasts produced by the method under test are equal to those expected to be correct by the naive method.

IV. APPLICATION TO MONSOON RAINFALL SERIES

The monsoon rainfall series are known to be homogeneous, Gaussion distributed and persistence free. Seasonality is not relevant here because we consider only particular season rainfall series. To remove the non-stationarity of trend type from the original data, Box and Jenkins (1976) suggested the repeated differencing and then fitting of the ARMA models to the differenced series. However, for the series of physical sciences like meteorology this operation is inapropriate (Katz and Skaggs, 1981).

As an initial step, for checking the stationarity, we plot the time series for visual inspection (Fig.2). Cramer's and Bartlett's Tests (WMO, 1971) are then applied for checking the constancy of the first 2 moments over the variable periods. Samples of size 50 are progressively selected from the time series and tested for the stability of mean and variance. Auto-correlograms and the Kendall's τ (Kendall, 1966) are also examined. After these tests, a latest stationary period of 1919-1968 is selected for model building such that an independent sample of moderate size is available for validation. Certain sample statistics alongwith Q-statistics of three rainfall series for this stationary epoch are given in Table 1.

Highly significant values of 'Q' indicate that the null hypothesis of random model should be rejected. Some ARMA model is expected to fit the series.

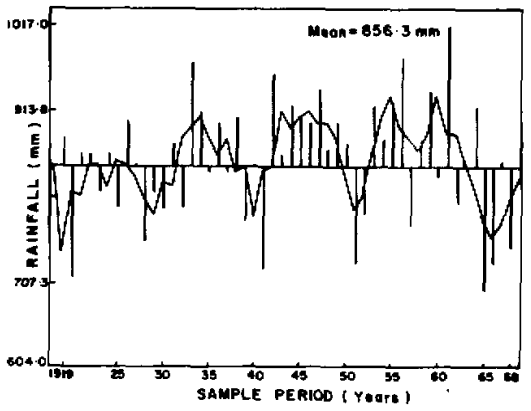

Fig.2. Original (Bars) and 3-point smoothed (curve) time series of India for Stationary period 1919–1968.

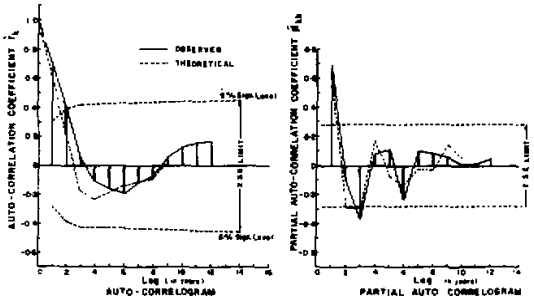

Fig.3. Sample ACF & PACF for 3-point smoothed series of India as a whole based on stationary period 1919–1968.

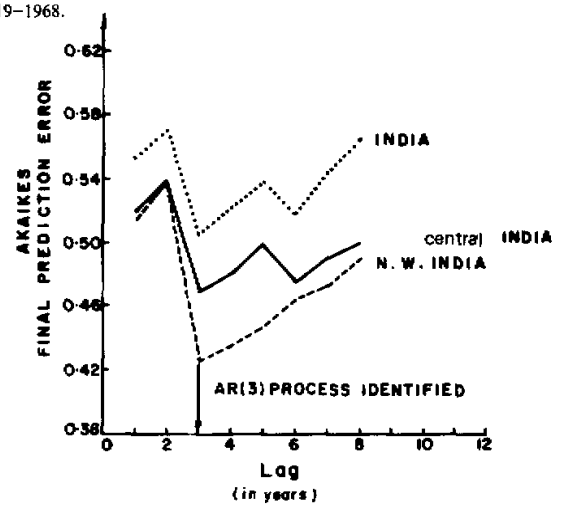

Fig.4. Akaikes final prediction error for selection of AR process for India, N.W. India and Central India.

The ACF and the PACF's for the all India rainfall series are presented in Fig.3. In general, all time series show significant lag-1 autocorrelations being 0.74, 0.76 and 0.75 respectively for whole India, NW India and Central India series. The general pattern of variation of ACF and PACF suggests that a low order model like AR(2), AR(3), MA(2) or ARMA (1,1) can represent the underlying process satisfactorily. The ACF is like a damped sine wave and the PACF cuts off for lag 3 indicating that AR(3) is the appropriate process. Akaike's FPE's (Fig.4) support this quantitatively.

Table 1. Mean, Variance, *C.V.* and Box-Pierce Q Statistic for Stationary Epoch 1919-1968

Region	Mean cm	Variance cm^2	*C.V.*(%)	Q with d.f. = 10
India	85.6	19.95	5.2	38.1 * *
N.W.India	49.5	34.44	11.9	62.3 * *
Central India	87.5	55.04	8.5	43.4 * *

* * Significant at 1%level.

Therefore AR(3) model is fitted to all time series. Initial estimates of AR(3) are obtained by Yule-Walker method which are further refined by Marquardt's algorithm (Pankratz, 1983) to obtain sufficient estimates. The estiamted parameters, innovation variance and variance explained etc. are given in Table 2.

Table 2. Model Parameters, Variance Explained (V.E.) and Innovation Variance for India, N.W. India and Central India

Region	Estimates Initial	sufficient	Mean Level μ Cms.	Innovation variance $\sigma^2 a$ Sq.cm
India	φ 1 = 0.714	0.8289		
	φ 2 = 0.194	0.2193	85.7	7.73
	φ 3 = -0.385	-0.4998		
V.E.	55%	61%		
N.W.India	φ 1 = 0.737	0.8432		
	φ 2 = 0.266	0.3365	49.4	10.15
	φ 3 = -0.475	-0.6121		
V.E.	62%	70%		
Central India	φ 1 = 0.723	0.8061		
	φ 2 = 0.244	0.2782	87.5	20.83
	φ 3 = -0.393	-0.4732		
V.E.	58%	62%		

The model equations can be written as;

For India,

$$Z_i = 85.7 + 0.8289(Z_{i-1} - 85.7) + 0.2193(Z_{i-2} - 85.7) - 0.4998(Z_{i-3} - 85.7) + ai\ , \quad ai \sim N\ (0,7.7)\ .$$

For N.W. India,

$$Z_t = 49.4 + 0.8432(Z_{t-1} - 49.4) + 0.3365(Z_{t-2} - 49.4) - 0.6121(Z_{t-3} - 49.4) + ai\ , \quad ai \sim N\ (0,10.1)\ .$$

For Central India,

$$Z_t = 87.5 + 0.8061(Z_{t-1} - 87.5) + 0.2782(Z_{t-2} - 87.5) - 0.4732(Z_{t-3} - 87.5) + ai\ , ai \sim N\ (0,20.8)\ .$$

To check how well the models fit sample for all regions, the residuals or innovations are calculated as $a(t) = Z(t) - \hat{Z}(t)$ and Ljung-Box Q^* (Pankratz, 1983) is computed. Correlation coefficient, Theil's U and RMSE are also evaluated. Results are shown in Table 3.

Table 3. Goodness of Fit Statistics for India, N.W. India and Central India

Region	Ljung-Box Q^* d.f. = 7	Corr.Coeff. between observed and modelled	Theil's U	R.M.S. Error (Cms)
India	7.20	0.82	0.76	2.42
N.W.India	6.43	0.86	0.68	2.82
Central India	11.44	0.83	0.77	3.96

All Q^* statistics are insignificant which implies that the model is adequate, $U < 1$ indicates that the model resembles the realization better than exponentially weighted moving average model. For checking the quality of parameters, intercorrelations are computed.

Table 4. Intercorrelations between AR(3) Parameters (Including Mean) for 3 Regions of India

	India				N.W. India				Central India			
	φ_1	φ_2	φ_3	mean	φ_1	φ_2	φ_3	mean	φ_1	φ_2	φ_3	mean
φ_1	1.00	−.69	.19	.12	1.00	−.71	.26	.11	1.00	−.69	.18	.14
φ_2		1.00	−.64	−.09		1.00	−.69	−.11		1.00	−.65	−.13
φ_3			1.00	−.06			1.00	−.01			1.00	−.01
mean				1.00				1.00				1.00

The intercorrelations are poor, (Table 4) implying that the parameters are independent of this particular realization. We can therefore, use this model for extrapolating future values. One should suspect that the estimates are somewhat unstable when absolute correlation coefficient between any two (including mean) estimated ARMA parameters is >0.9 (Pankratz, 1983). In our case the intercorrelations are well below 0.9, so the estimated model is satisfactory. The parameters satisfy the condition of stationarity and the parameters φ_1 and φ_3 are significant.

We have used this fitted model to make forecasts during an independent period of 18 years (1970-1987). One step ahead forecast and their 95% confidence intervals are shown in Figure 5. Actual forecasts computed from smoothed series are tabulated in Table 5. The Heidke skill scores for 3 categories forecasts are given in the bottom of the same Table. Bracketed scores are for smoothed series.

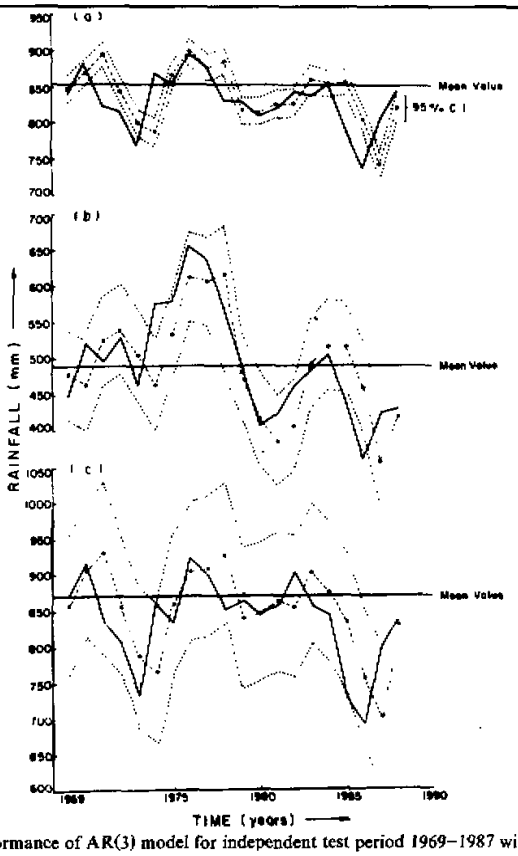

Fig.5. Performance of AR(3) model for independent test period 1969–1987 with 95% confidence intervals of forecasts for a) India as a whole b) N.W. India, c) Central India. —— Observed Sm–3 Series • ---- • Forecasted through AR (3) process ::::::: } 95%CI

V. DISCUSSION AND CONCLUSION

The autoregressive model building procedures are employed to characterize smoothed rainfall series of India and its two major regions. Selected samples are stationary and thus appropriate for UBJ analysis. The AR(3) model is selected on the basis of ACF, PACF patterns, Akaikes FPE criterion and minimum residual variance. The model adequacy is tested through diagnostic checking. Further the quality of model like closeness of fit, statistical significance of parameters, intercorrelations and stationarity of modelled process [AR(3)] is tested. Model potential is evaluated on independent test data set.

The original rainfall time series does not show significant autocorrelation or partial autocorrelation at any lag, so it lacks the significant structure and hence cannot be fruitfully modelled by AR class of models. However, after smoothing some significant structure emerges which is modelled here and used for linear extrapolation. Instead of assuming stationarity or removing non–stationarity by nonseasonal differencing, we have selected a stationary part of the 3–point moving averaged smoothed series for identifying a proper stochastic process. It is certainly true that climatological series like Indian rainfall cannot be explained entirely by its own past values without considering any effect from other meteoro-

logical parameters. However, interannual variation to certain extent can be accounted for by the simple class of univariate ARMA models. AR(3) model is suggested to be the most appropriate model in the above analysis. Each monsoon event is final output of various non-linear interactions between regional and planetary scale parameters and generally most recent information could be more useful for its prediction. Many researchers (Parthasarathy et al., 1988; Gowariker et al., 1989; Thapliyal, 1990) have developed forecasting methods based on multiple regression with different parameters. Generally the lead time of such forecasts is a few months. The present method eventhough explains small variance of the original series, it provides the forecast with lead time of nearly a year. Univariate linear extrapolation method has limited skill as it explains only the feedback mechanism part of the monsoon process on regional scale. However the skill can be improved either by combining with some independent statistical schemes explaining teleconnections with Indian monsoon or considering non-linear extrapolation methods like Threshold Auto Regression (TAR) or multivariate time series models etc.. It is hoped that these forecasts can be considered along with the forecasts prepared by other methods, for improving the lead time and skills of forecasts.

Table 5. Actual and Forecasted Values (Cms.) for Independent Test Data Period: 1970–1987 and Heidke Skill Score

Year	India		N.W.India		Central India	
	actual	*forecast*	*actual*	*forecast*	*actual*	*forecast*
1970	93.9	96.3	53.6	61.9	104.4	100.8
1971	88.6	84.5	59.6	40.8	84.8	81.6
1972	65.3	86.8	36.4	43.6	61.3	90.5
1973	91.2	99.7	62.3	64.7	96.2	111.4
1974	74.7	84.6	40.1	52.5	63.8	78.6
1975	96.0	71.6	71.3	40.0	98.7	69.3
1976	85.5	89.9	62.4	48.7	88.3	95.6
1977	88.1	89.2	63.4	50.1	90.4	84.7
1978	90.8	90.5	65.8	56.2	92.4	94.3
1979	70.8	87.5	34.4	56.2	73.2	96.4
1980	88.3	84.4	46.4	45.4	93.5	86.6
1981	85.2	87.2	41.4	44.1	87.2	87.3
1982	73.5	75.2	38.7	27.6	76.0	77.0
1983	95.5	90.8	58.2	41.0	107.5	93.8
1984	83.5	90.1	48.2	51.2	74.1	88.1
1985	78.7	78.5	44.8	49.0	71.6	81.3
1986	74.7	95.6	39.8	61.9	72.7	104.3
1987	68.8	88.9	23.2	52.8	62.9	82.2
Heidke skill score	0.345 (0.486)		0.029 (0.408)		0.327 (0.284)	

The following are the main conclusions of the present study.

1) The monsoon rainfall series smoothed by 3-point running mean for India, N.W. India and Central India have a stationary period of 1919–1968.

2) 3-point smoothed series of India and its tow major regions are adequately represented by AR(3) process.

3) Variance explained by AR(3) models of 3-point smoothed series are satisfactory, be-

ing 61% for India, 70% for N.W. India and 62% for Central India.

4) The AR(3) model could be used as a first guess to forecast categoriwise, viz. Excess, normal and deficient rainfall with nearly one year lead time.

This paper has examined the monsoon rainfall time series for the period 1871–1989. Recently some authors have constructed the all India rainfall series by using limited stations data as far back as 1813. It remains to be examined if similar characteristics as reported in this study hold good for the rainfall series of the last century.

The authors wish to express their sincere thanks to the Director, I.I.T.M. for facilities provided and to Dr. S.S. Singh for encouragement. Thanks are also due to Dr. B. Parthasarathy, Dr. Nityanand Singh and Anonymous referees for making fruitful suggestions.

REFERENCES

Akaike H. (1971), Auto-Regressive Model fitting for control, *Ann. Inst. Stat. Maths.*, **23:** 163–180.

Box G.E.P. and Jenkins G.M. (1976), *Time Series Analysis Forecasting* and *Control*, Holden-Day, Oakland, 57 pp.

Brown T. J. and Fluek J. A. (1987), *Exploratory time series modelling and prediction of equatorial Pacific warming and cooling events Tenth Conference on Prob. and Statistics in Atmospheric Sciences*, Canada, *A. M. S.*, 178–182.

Chu P.S. and Katz R.W. (1985), Modelling and Forecasting the Southern oscillations- A time domain approach, *Mon. Wea. Rev.*, **113:** 1876–1888.

Elliot W. P. & Angell J. K. (1987), The relation between Indian Monsoon Rainfall, the southern oscillation and Hemispheric air sea temperature, 1884–1984., *J. Clim. Appl. Meteol.*, **26:** 943–948.

Gowariker V. V., Thapliyal, V., R. P. Sarker, G. S. Mandal and D. R. Sikka (1989), Parametric and power regression models-New approach to long range forecasting, *Mausam*, **40:** 115–122.

Jolliffe I. T. (1983), Quasi-periodic meteorological series and second order auto-regressive processes, *J. of Climatology*, **3:** 413–417.

Katz R. W. and Skaggs R. H. (1981), On use of Autoregressive-Moving average process to the meteorological time series, *Mon. Wea. Rev.*, **109:** 479–484.

Kendall, M. G and A. Stuart (1966), *The advanced theory of statistics*, Vol.3, Griffin, London, 585 pp.

Normand C. (1953), Monsoon seasonal forecasting, *Quar. J. Roy. Met. Soc.*, **79:** 463–473.

Pankratz Alan (1983), *Forecasting with univariate Box-Jeckins models concept and cases*, John Wiley, New York, 562 pp.

Pant, G.B., Rupa Kumar K. Parthasarathy B. and Borgaonkar H.P. (1988), Long term variability of the Indian Summer Monsoon and related parameters, *Adv. Atmos. Sci.* **5(4):** 469–481.

Parthasarathy, B., Sontakke N. A., Munot A. A., Kothawale, D. R. (1987), Drought / Floods in the Summer Monsoon Season over different meteorological subdivisions of India for the period 1871–1984, *J. of Climatology*, **7:** 57–70.

------(1990), Vagaries of Indian Monsoon and its relationships with regional / global circulations, *Mausam*, **41:** 301–308.

Thapliyal V. (1982), Stochastic dynamic model for long range prediction of monsoon rainfall in Peninsular India, *Mausam*, **33:** 399–404.

-------(1986), Long Range Forecasting of Monsoon Rainfall in India, *LRF Research Series 6 (III)*, WMO / TD **87:** 723–732.

-------(1990), Long Range Prediction of summer monsoon rainfall over India: Evolution and development of models. *Mausam*, **41:** 339–346.

Theil, H. (1966), *Applied Economic Forecasting*, Amsterdam: North Holland Publishing Co.: 26–32.

WMO (1971), *Climate Change, Tech. Note No.79 (annex III)* Geneva: 63–69.

Yao C. S. (1983), Fitting a linear Auto-regressive model for long range forecasting, *Mon. Wea. Rev.*, **111:** 692–700.